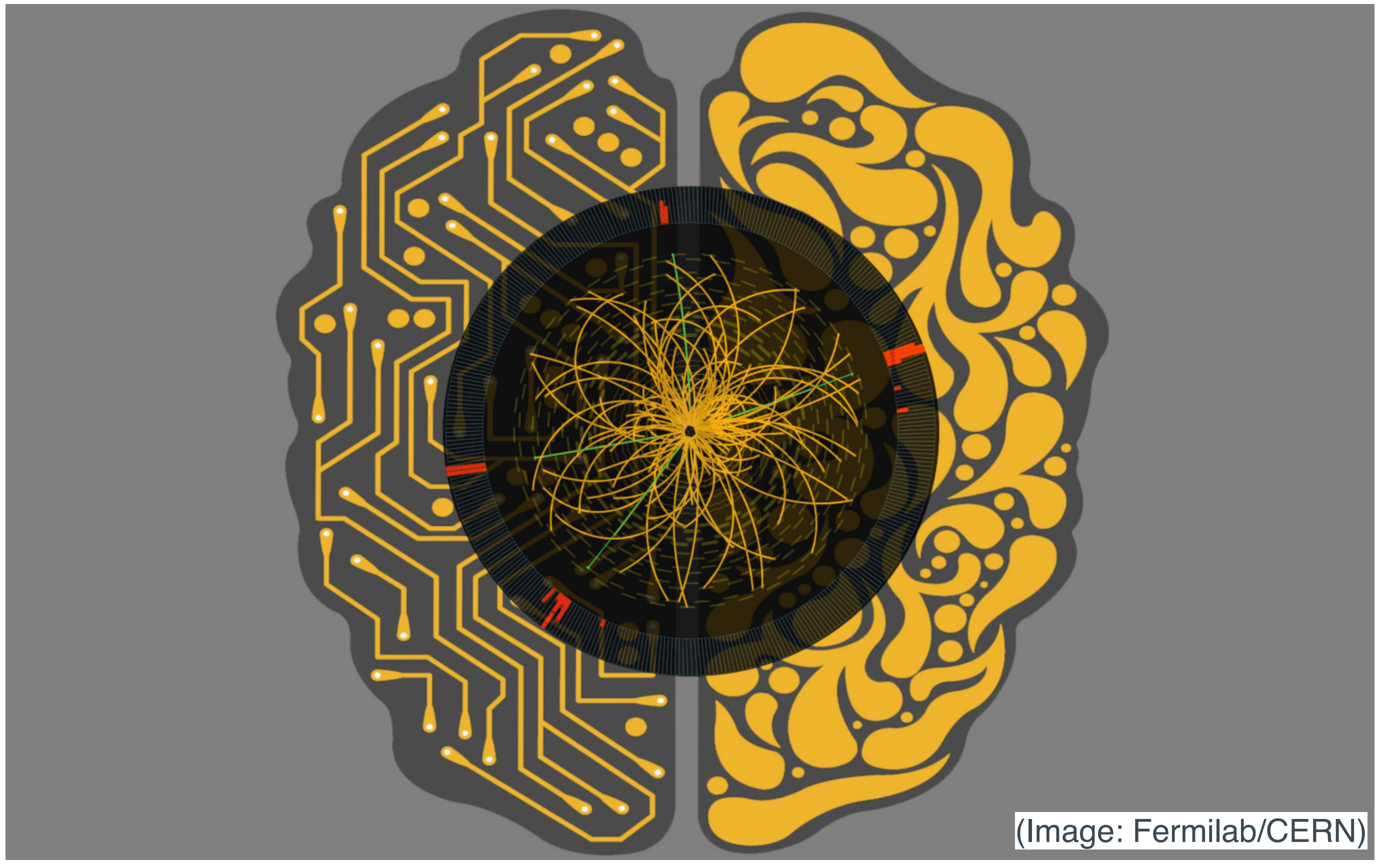
PHY 835: Collider Physics Phenomenology

Machine Learning in Fundamental Physics

Gary Shiu, UW-Madison



Lecture 15: Variational methods & Mean Field Theory

Recap of Lecture 14

- K-means clustering
- Agglomerative clustering
- Density-based (DB) clustering
- Gaussian mixture models

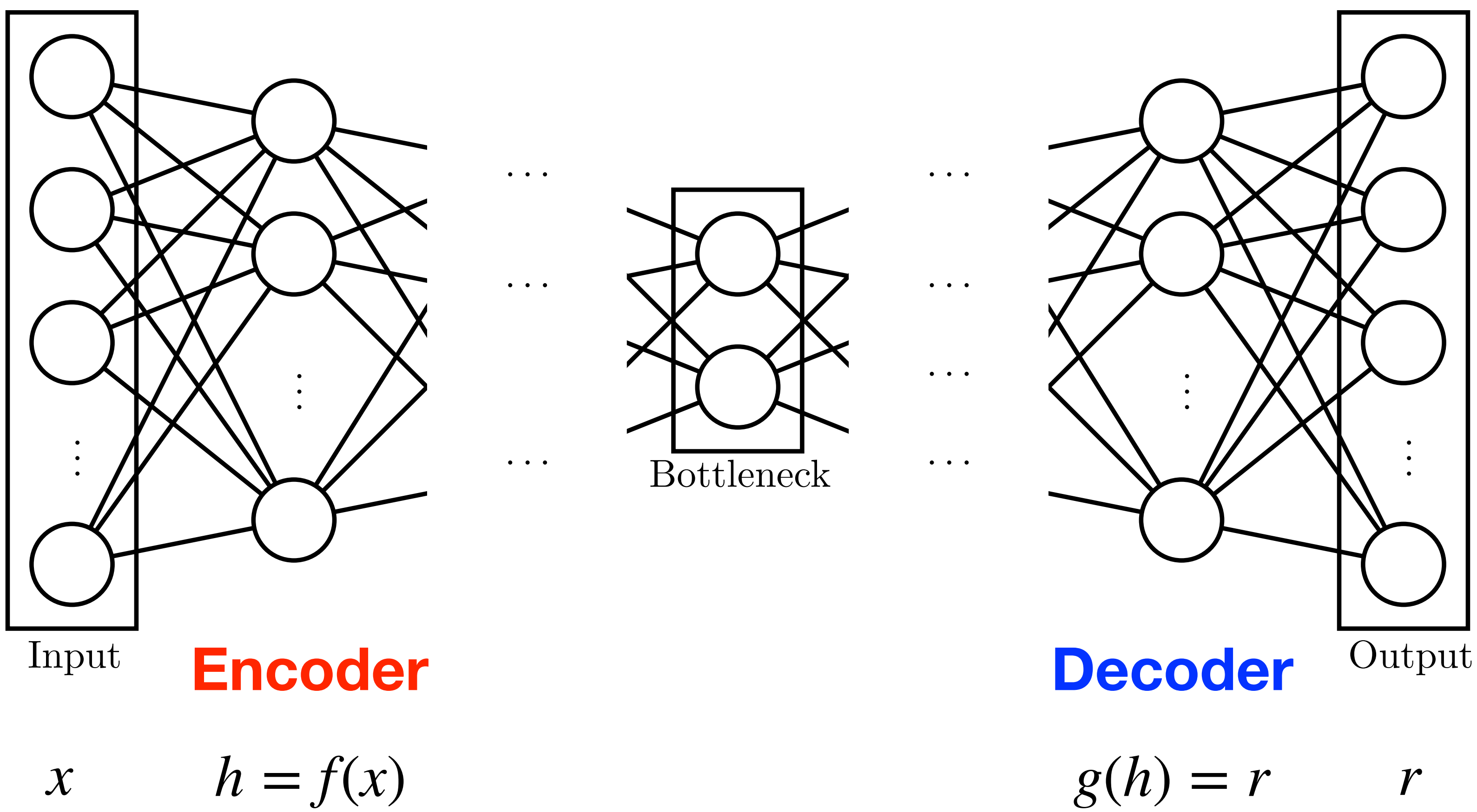
Outline for today

- Autoencoder
- Variational methods and Mean Field Theory (MFT)
- Expectation-Maximization (EM)

References: 1803.08823, Deep Learning Book
https://blog.keras.io/building-autoencoders-in-keras.html

Autoencoder

- Autoencoder: Copy input to its output via bottleneck



$$g(f(x)) = r$$

$$C = \sum_i (x_i - r_i)^2$$

Aim at reproducing input: can be trained as a neural network

- Very similar to PCA, but here: encoder and decoder can be non-linear functions.

Autoencoder

- Non-linear function makes this more powerful: can copy everything in principle, but usually not in practice.

- Ways around over-fitting:

  - Regularization

  - Encourage model to have further properties (e.g. variational autoencoder)

- Examples: Building an auto encoder for Ising and MNIST datasets: Notebooks 19 and 20:

https://physics.bu.edu/~pankajm/ML-Notebooks/HTML/NB19_CXVII-Keras_VAE_MNIST.html

https://physics.bu.edu/~pankajm/ML-Notebooks/HTML/NB20_CXVII-Keras_VAE_ising.html

Relative vs Absolute Probabilities

- Need to accurately represent the underlying probability distribution.

- Much easier to learn relative weights than absolute probabilities.

$$\frac{p(\mathbf{x}_1)}{p(\mathbf{x}_2)} = \mathrm{e}^{-\beta(E(\mathbf{x}_1)-E(\mathbf{x}_2))} \qquad \text{vs} \qquad p(\mathbf{x}) = \frac{\mathrm{e}^{-\beta E(\mathbf{x})}}{Z_p} \quad \text{with} \quad Z_p = \mathrm{Tr}_{\mathbf{x}}\mathrm{e}^{-\beta E(\mathbf{x})}$$

where $\beta =$ inverse temp. and $E(\mathbf{x}, \theta) =$ energy of state $\mathbf{x}$.

- The partition function Z_p is computationally intractable, e.g., the Ising model with N binary spins, trace involves summing 2^N terms.

- Monte-Carlo based methods to draw samples from the underlying distribution and use these samples to estimate Z_p, e.g., Markov Chain Monte Carlo (MCMC) and annealed importance sampling.

Variational Methods

- Approximate the probability distribution $p(\mathbf{x})$ and partition function by a **variational distribution** $q(\mathbf{x}, \theta_q)$ whose Z_p can be calculated exactly; θ_q is chosen to make $q(\mathbf{x}, \theta) \approx p(\mathbf{x})$ as much as possible.

- **Mean-Field Theory (MFT):** factorized distribution.

- **Expectation-Maximization (EM):** not only for GMM but a general variational method for latent (hidden) variables.

- First illustrate the idea of variational MFT with the Ising Model, and show how MFT can be formulated as an EM problem.

- Optimizing the approximate probability distribution amounts to minimizing the KL divergence $D_{KL}(q\,||\,p)$.

Ising Model

- The energy of a given spin configuration is given by the Hamiltonian:

$$E(\boldsymbol{s}, \boldsymbol{J}) = -\frac{1}{2} \sum_{i,j} J_{ij} s_i s_j - \sum_i h_i s_i,$$

- (J_{ij}, h_i) are typically uniform, or in the case of disordered systems, drawn from some probability distribution (quenched disorder).

- Probability of finding the system in a given spin configuration:

$$p(\boldsymbol{s}|\beta, \boldsymbol{J}) = \frac{1}{Z_p(\boldsymbol{J})} \mathrm{e}^{-\beta E(\boldsymbol{s}, \boldsymbol{J})},$$

$$Z_p(\beta, \boldsymbol{J}) = \sum_{\{s_i = \pm 1\}} \mathrm{e}^{-\beta E(\boldsymbol{s}, \boldsymbol{J})},$$

- Subscript p of the partition function $Z_p(\beta, \mathbf{J})$ corresponds to the probability distribution $p(\mathbf{s} \,|\, \beta, J)$.

Ising Model

- In general not possible to evaluate the sum of 2^N terms of $Z_p(\beta, \mathbf{J})$ in closed form ⇒ represents challenges for extracting physics e.g.

Free energy: $$\beta F_p(\boldsymbol{J}) = -\log Z_p(\beta, \boldsymbol{J}) = \beta \langle E(\boldsymbol{s}, \boldsymbol{J}) \rangle_p - H_p,$$

with

Entropy: $$H_p = -\sum_{\{s_i = \pm 1\}} p(\boldsymbol{s}|\beta, \boldsymbol{J}) \log p(\boldsymbol{s}|\beta, \boldsymbol{J})$$

- **Idea:** approximate $p(\mathbf{s} \,|\, \beta, \mathbf{J})$ by a variational distribution $q(\mathbf{s}, \theta)$ and vary θ to make the two distributions as closed as possible.

- **Variational free energy:**

$$\beta F_q(\boldsymbol{\theta}, \boldsymbol{J}) = \beta \langle E(\boldsymbol{s}, \boldsymbol{J}) \rangle_q - H_q,$$

Ising Model

- Recall that the KL divergence has the following properties:
  - **Positivity:** $D_{KL}(p\,||\,q) \geq 0$ with equality iff $p = q$.
  - **Asymmetry:** $D_{KL}(p\,||\,q) \neq D_{KL}(q\,||\,p)$.
- Positivity implies that $F_q(\mathbf{J}, \theta) \geq F_p(\mathbf{J}, \theta)$ with equality iff $q = p$ (in the sense of distribution); best variational free energy minimizes $D_{KL}(q\,||\,p)$.
- In MFT, $q(\mathbf{s}, \theta)$ is taken to be a factorized distribution:

$$q(\boldsymbol{s}, \boldsymbol{\theta}) = \frac{1}{Z_q} \exp\left(\sum_i \theta_i s_i\right) = \prod_i \frac{\mathrm{e}^{\theta_i s_i}}{2\cosh\theta_i}$$

- This simplification enables closed form expressions. Drawback is ignoring correlations between spins (less important for large N).

Ising Model

- With the MFT ansatz, the entropy H_q of the distribution q:

$$H_q(\boldsymbol{\theta}) = -\sum_{\{s_i=\pm 1\}} q(\boldsymbol{s},\boldsymbol{\theta}) \log q(\boldsymbol{s},\boldsymbol{\theta})$$
$$= -\sum_i q_i \log q_i + (1-q_i)\log(1-q_i), \qquad \text{where } q_i = \frac{e^{\theta_i}}{2\cosh\theta_i}$$

- The mean value of s_i (on-site magnetization):

$$m_i = \langle s_i \rangle_q = \sum_{s_i=\pm 1} s_i \frac{e^{\theta_i s_i}}{2\cosh\theta_i} = \tanh(\theta_i).$$

- Because the spins are independent, the average energy is simple:

$$\langle E(\boldsymbol{s},\boldsymbol{J})\rangle_q = -\frac{1}{2}\sum_{i,j} J_{ij} m_i m_j - \sum_i h_i m_i.$$

- The total variational free-energy: $\beta F_q(\boldsymbol{J},\boldsymbol{\theta}) = \beta\langle E(\boldsymbol{s},\boldsymbol{J})\rangle_q - H_q$

Ising Model

- Minimizing the variational free-energy with respect to θ:

$$0 = \frac{\partial}{\partial \theta_i} \beta F_q(\boldsymbol{J}, \boldsymbol{\theta}) = 2 \frac{dq_i}{d\theta_i} \left(-\beta \left[\sum_j J_{ij} m_j + h_i \right] + \theta_i \right) \quad \Rightarrow \quad \theta_i = \beta \sum_j J_{ij} m_j(\theta_j) + h_i.$$

- For uniform couplings, $h_i = h$ and $J_{ij} = J \Rightarrow \theta_i = \theta$ by symmetry &

$m = \tanh(\theta)$ and $\theta = \beta(zJm(\theta) + h)$, where z is the coordination number of the lattice (i.e. the number of nearest neighbors)

- The MFT of Ising Model can be formulated as an EM procedure, similar to GMM and K-means clustering discussed earlier:

1. *Expectation*: Given a set of assignments at iteration t, $\boldsymbol{\theta}^{(t)}$, calculate the corresponding magnetizations $\boldsymbol{m}^{(t)}$ using Eq. (167)
2. *Maximization*: Given a set of magnetizations m_t, find new assignments $\theta^{(t+1)}$ which minimize the variational free energy F_q. From, Eq. (170) this is just

$$\theta_i^{(t+1)} = \beta \sum_j J_{ij} m_j^{(t)} + h_i.$$

Drawbacks of MFT

- Cannot capture correlations between the spins, leading to
  - Wrong value of the critical temperature for the 2D Ising Model.
  - Erroneously predicts the existence of a phase transition in one dimension at a non-zero temperature.
- Despite these drawbacks, MFT often yield qualitatively and even quantitatively precise predictions (especially in high dimensions).
- Illustrate the general relation between variational methods and the EM procedure.

Expectation-Maximization (EM)

- Variational MFT has been developed to perform MLE. Its close relationship with EM was worked out in Neal and Hinton (1998).

- Latent variables make MLE difficult to implement. EM gets around this difficulty by using an iterative two-step procedure.

- Let $\mathbf{x} =$ set of visible variables, $\mathbf{z} =$ set of latent variables, $p(\mathbf{z}, \mathbf{x} \,|\, \theta)$= probability distribution from which $\mathbf{x}$ and $\mathbf{z}$ are drawn.

- Since we can only observe $\mathbf{x}$, we wish to find the parameters θ that maximizes the probability of the **observed data**:

$$L(\boldsymbol{\theta}) = \langle \log p(\boldsymbol{x}|\boldsymbol{\theta}) \rangle_{P_{\mathbf{x}}}$$ **log likelihood**

Expectation-Maximization (EM)

- Initialize $\theta^{(0)}$ and iterating the variational parameters $\theta^{(t)}, t = 1,2,\ldots$

1. **Expectation step (E step):** Given the known values of observed variable $\boldsymbol{x}$ and the current estimate of parameter $\boldsymbol{\theta}_{t-1}$, find the probability distribution of the latent variable $\boldsymbol{z}$:

$$q_{t-1}(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{x})$$

2. **Maximization step (M step):** Re-estimate the parameter $\boldsymbol{\theta}^{(t)}$ to be those with maximum likelihood, assuming $q_{t-1}(\boldsymbol{z})$ found in the previous step is the true distribution of hidden variable $\boldsymbol{z}$:

$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta}) \rangle_{q_{t-1}}$$

- EM iteration increases the true log-likelihood $L(\theta)$, or at worst leaves it unchanged. In most models, this iteration procedure converges to a local maximum of $L(\theta)$.

- With data $\mathbf{z}$ missing, we cannot just maximize $L(\theta)$ directly since parameter θ might couple both $\mathbf{z}$ and $\mathbf{x}$.

Expectation-Maximization (EM)

- Idea: optimizing another objective function, $F_q(\theta)$, constructed based on estimates of the hidden variable distribution $q(\mathbf{z}\,|\,\mathbf{x})$.

- This objective function is the **variational free energy**:

$$F_q(\boldsymbol{\theta}) := -\langle \log p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})\rangle_{q,P_{\mathbf{x}}} - \langle H_q \rangle_{P_{\mathbf{x}}}$$

- The **true free-energy** is:

$$-F_p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) = \langle \log p(\boldsymbol{x}|\boldsymbol{\theta})\rangle_{P_{\mathbf{x}}}.$$

- This minus sign is chosen because the free-energy is minus log of the partition function, often omitted in the ML literature (be cautious).

Expectation-Maximization (EM)

- Minimizing the difference $F_q(\boldsymbol{\theta}) - F_p(\boldsymbol{\theta}) = \langle f_q(\boldsymbol{x}, \boldsymbol{\theta}) - f_p(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{P_{\mathbf{x}}}$,

where
$$f_q(\boldsymbol{x}, \boldsymbol{\theta}) - f_p(\boldsymbol{x}, \boldsymbol{\theta}) = \log p(\boldsymbol{x}|\boldsymbol{\theta}) - \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta}) \; + \; \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log q(\boldsymbol{z}|\boldsymbol{x})$$
$$= \overset{1}{\overset{\parallel}{\sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x})}} \log p(\boldsymbol{x}|\boldsymbol{\theta}) - \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta}) \; + \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log q(\boldsymbol{z}|\boldsymbol{x})$$
$$= -\sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta})} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log q(\boldsymbol{z}|\boldsymbol{x})$$

Note typo in 1803.08823

Using Bayes' theorem

$p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta}) / p(\boldsymbol{x}|\boldsymbol{\theta})$

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})}$$
$$= D_{KL}(q(\boldsymbol{z}|\boldsymbol{x}) || p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})) \geq 0$$

- We thus prove our earlier assertion that the difference between the approximate and the true distributions is the KL divergence.

- The variational free energy is an upper bound for true free-energy.

Expectation-Maximization (EM)

- Since H_q does not depend on θ, the M-step is equivalent to minimizing the variational free-energy $F_q(\theta)$.

- Less obvious is that the E-step can also viewed as the optimization of this variational free-energy. It turns out:

$$q_{t-1}(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{x})$$

is the unique probability that minimizes $F_q(\theta)$ (now seen as a functional of q). Hint: taking the functional derivative of $F_q(\theta)$ plus a Lagrange multiplier λ (that enforces $\sum_z q(z) = 1$) w.r.t. $q(z)$:

$$-\log p(\mathbf{z}\,|\,\theta, \mathbf{x}) + \log q(\mathbf{z}) + 1 + \lambda = 0 \Rightarrow q(\mathbf{z}) \propto p(\mathbf{z}\,|\,\theta, \mathbf{x})$$

- The normalization condition enforced by the Lagrange multiplier implies $q(\mathbf{z}) = p(\mathbf{z}\,|\,\theta, \mathbf{x})$. More details in Neal and Hinton (1998).

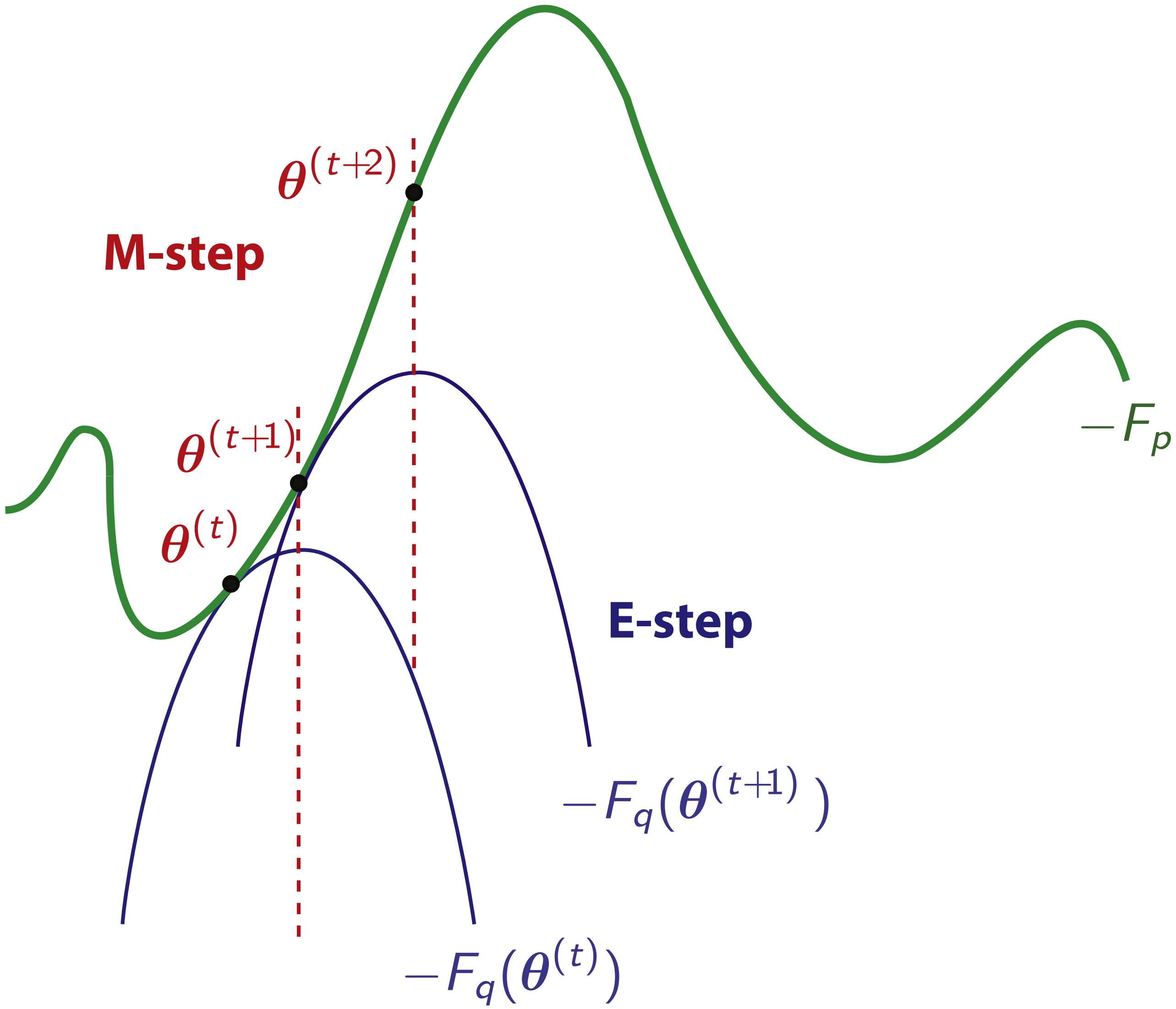
Expectation-Maximization (EM)

1. *Expectation step:* Construct the approximating probability distribution of unobserved $\boldsymbol{z}$ given the values of observed variable $\boldsymbol{x}$ and parameter estimate $\boldsymbol{\theta}^{(t-1)}$:

$$q_{t-1}(\boldsymbol{z}) = \arg\min_q F_q(\boldsymbol{\theta}^{(t-1)})$$

2. *Maximization step:* Fix q, update the variational parameters:

$$\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} -F_{q_{t-1}}(\boldsymbol{\theta}).$$



Expectation-Maximization (EM)

Table 1
Analogy between quantities in statistical physics and variational EM.

Statistical physics	Variational EM
Spins/d.o.f.: $\boldsymbol{s}$	Hidden/latent variables $\boldsymbol{z}$
Couplings /quenched disorder: $\boldsymbol{J}$	Data observations: $\boldsymbol{x}$
Boltzmann factor $e^{-\beta E(\boldsymbol{s},\boldsymbol{J})}$	Complete probability: $p(\boldsymbol{x}, \boldsymbol{z}\|\boldsymbol{\theta})$
Partition function: $Z(\boldsymbol{J})$	Marginal likelihood $p(\boldsymbol{x}\|\theta)$
Energy: $\beta E(\boldsymbol{s}, \boldsymbol{J})$	Negative log-complete data likelihood: $-\log p(\boldsymbol{x}, \boldsymbol{z}\|\boldsymbol{\theta}, m)$
Free energy: $\beta F_p(\boldsymbol{J}\|\beta)$	Negative log-marginal likelihood: $-\log p(\boldsymbol{x}\|m)$
Variational distribution: $q(\boldsymbol{s})$	Variational distribution: $q(\boldsymbol{z}\|\boldsymbol{x})$
Variational free-energy: $F_q(\boldsymbol{J}, \boldsymbol{\theta})$	Variational free-energy: $F_q(\boldsymbol{\theta})$

Experiment with the python notebook:

https://physics.bu.edu/~pankajm/ML-Notebooks/HTML/NB16_CXIII-EM_coin_toss.html

Summary

- Autoencoder
- Variational methods and Mean Field Theory (MFT)
- Expectation-Maximization (EM)